Модель обмена знаниями в системах гуманитарных исследований

И. В. Кравцов, К. А. Багимова

Петрозаводский государственный университет, Петрозаводск, Россия

Ижевский государственный технический университет, Ижевск, Россия

На наш взгляд необходимо переосмыслить понимание цифровой публикации каких-либо текстов, являющихся памятниками культурно-исторического наследия. Кажущийся когда-то новаторским подход переноса публикации из печатной в цифровую форму в виде изображений или транскрипций текстов сейчас удобен лишь как средство упрощенного удаленного доступа. В большинстве случаев не используются обширные возможности, которые предоставляют современные информационные технологии [Кравцов и др. 2007].

Даже если публикация готовится в специализированной информационной системе, такой, как, например, «Манускрипт» (URL: http://manuscripts.ru/), то дополнительные возможности по работе со структурой и семантикой текстов возможны лишь в рамках этой системы, в большинстве случаев в локальном доступе. Извлечение текстов и информации о них из таких систем осложнено особенностями внутреннего формата хранения. Можно сказать, тексты являются полноправной частью системы, электронной библиотеки и не могут быть отчуждаемы.

Представление информационных единиц в сети как отдельных объектов с уникальным идентификатором является одним из принципов формирования Семантического веба – интернет-пространства нового поколения. Вторым принципом служит необходимость оформления этих информационных объектов в виде «двойного» текста: с одной стороны, это информационное сообщение для человека, с другой – информация для компьютера. Таким образом, эта информация о структуре и семантике объекта становится машиночитаемой и может использоваться более продвинутыми сервисами поиска и анализа в отличие от существующих на данный момент [Баранов и др. 2007].

Авторами статьи ведется работа над созданием абстрактной модели описания структуры и семантики текстов документов письменного культурного

наследия [Кравцов 2008]. Отражение модели в определенном формате данных будет являться, с одной стороны, способом создания информационных объектов в семантическом вебе, а с другой стороны, будет являться инвариантом, форматом обмена текстами и знаниями о текстах между различными информационными системами.

Необходимость создания подобного формата обмена назрела уже давно и была озвучена на прошлой конференции «Современные информационные технологии и письменное наследие» [Современные 2006]. В целом, потребность разработчиков и публикаторов в консолидации вылилась в формирование сообщества «Письменное наследие» [Баранов и др. 2007], одной из целей которого является переход в единое информационное пространство как на уровне межличностного взаимодействия, так и на уровне разделения между участниками оцифрованных текстовых ресурсов и инструментов работы с ними.

Задачи анализа текстов, например, задачи исследования рукописных исторических документов, могут быть поставлены с точки зрения источниковедения, истории, лингвистики. В указанных, а также во многих других гуманитарных дисциплинах довольно схожи методики анализа текстовых источников. Все они, так или иначе, пытаются формализовать текст, выделить необходимые категории, построить на уровне абстракции свои модели, провести с построенными моделями характерные исследования и попытаться интерпретировать результат.

Для таких задач можно предложить универсальную модель описания формализованного текста и извлеченных из него знаний. Универсальная модель должна удовлетворять следующим требованиям:

- выделять произвольные единицы текста как обособленные объекты;
- формировать связь произвольного числа объектов;
- позволять строить произвольные иерархии объектов и связей;
- соотносить как объекты, так и связи с произвольными смысловыми категориями;
- привязывать к объектам и связям различные показатели (числовые, номинальные, вероятностные и пр.);

- позволять переходить от моделей текстов к моделям более высокого уровня (например, к модели коллекций текстов).

Авторами статьи предлагается в качестве такой универсальной модели – модель «структурно-семантического пространства». Данное пространство состоит из набора измерений, количество которых потенциально бесконечно, и точек в этом пространстве. Каждое измерение является фиксированным набором значений, определяет некоторую шкалу. Каждая точка отражает факт взаимосвязи значений на фиксированном наборе измерений.

При работе с текстом существует понятие базового измерения – это отложенные на шкале слова, формирующие текст в порядке их появления. Практически любая точка в структурно-семантическом пространстве определяется хотя бы одним базовым измерением и некоторым набором других измерений. Например, для соотнесения элементов текста с определенными смысловыми категориями (выделения персоналий в тексте) создается точка в двумерном подпространстве, где одно измерение базовое, а другое – шкала категории объектов («правители, воеводы, бояре и проч.»). Для создания связи между двумя объектами текста берется два базовых измерения и, если необходимо, измерение категорий связей, и в этом подпространстве ставится точка. Соответственно, для создания N-арной связи берется N базовых измерений.

Таким образом, задача преобразования любой внутренней структуры хранения в обобщенную модель сводится к разложению в подобном пространстве. Кроме того, использование самих текстов в качестве базы пространства позволит связывать и дополнять проведенные с ними в разных информационных системах исследования. Текст – это только образующие его словоформы, каждая из которых имеет свой уникальный идентификатор. И, например, информация о разбиении его на строки и страницы – это уже дополнительная информация, знания, особенность конкретного источника или издания, отдельное измерение в структурно-семантическом пространстве.

Попробуем проиллюстрировать добавление различной информации о тексте на примере задачи подготовки цифрового издания полного собрания сочинений М. В. Ломоносова [О коллекции 2007]. Публикация осуществляется

в формате информационно-аналитической системы «Манускрипт» (URL: http://manuscripts.ru/), для этого бумажные источники сканируются и распознаются, распознанный текст вычитывается и размечается с помощью специализированных XML схем в трех вариантах: геометрической, лингвистической и структурно-функциональной иерархиях.

Посмотрим все три способа разметки на примере одного абзаца текста:

Присовокупление III

6. Если требуется превратить какое-либо твердое тело в жидкое то необходимо затруднить взаимное сцепление отдельных его частиц.

В геометрической разметке он будет выглядеть вот так:

```
<l no="25">  <wf id="w345">Присовокупление</wf><wf id="w346">III</wf></l>
<l no="26">  <wf id="w347">6</wf>.<wf id="w348">Если</wf><wf id="w349">требуется</wf> <wf id="w350">превратить</wf><wf id="w351">какое-либо</wf><wf id="w352">твердое</wf><wf id="w353">тело</wf></l>
<l no="27"><wf id="w354">в</wf><wf id="w355">жидкое</wf>,<wf id="w356">то</wf><wf id="w357">необходимо</wf><wf id="w358">затруднить</wf><wf id="w359">взаимное</wf><wf id="w360">сцепление</wf><wf id="w361">от<lb/></l>-
<l no="28">дельных</wf><wf id="w362">его </wf><wf id="w363">частиц</wf>.</l>
```

Видно, что каждой словоформе <wf> присвоен уникальный (в данном случае в пределах текста) идентификатор. Кроме того, тегом <l> выделены строки. В модели структурно-семантического пространства эту разметку можно представить как пространство из двух измерений: первое – измерение словоформ, второе – измерение строк. Точкой в пространстве или, другими словами, фактом будет указание на то, что определенная словоформа (идентификатор словоформы) попала в определенную строку (идентификатор строки). В точке можно в качестве значения хранить само значение словоформы и тогда при переносе слов будет два факта вхождения

словоформы в разные строки: в первом факте - начало словоформы, во втором факте - окончание словоформы (в примере слово «от-дельных»).

В данном случае идентификаторы словоформ упорядочены, и по ним можно восстановить порядок слов в строке. Если бы это было не так, то в пространстве пришлось бы добавить еще одно измерение порядка и указать для каждой словоформы как факт её позицию в пределах предложения или текста.

В лингвистической иерархии наш пример будет выглядеть вот так:

```
<div type="sent" no="32">
<wf id=”w345”>Присовокупление</wf><wf  id=”w346”>III</wf>
</div>
<div type="sent" no="33">
<wf id=”w347”>6</wf>.<wf id=”w348”>Если</wf><wf id=”w349”>требуется</wf>
<wf id=”w350”>превратить</wf><wf id=”w351”>какое-либо</wf>
<wf id=”w352”>твердое</wf><wf id=”w353”>тело</wf><wf id=”w354”>в</wf>
<wf id=”w355”>жидкое</wf>,<wf id=”w356”>то</wf>
<wf id=”w357”>необходимо</wf><wf id=”w358”>затруднить</wf>
<wf id=”w359”>взаимное</wf><wf id=”w360”>сцепление</wf>
<wf id=”w361”>отдельных</wf><wf id=”w362”>его </wf>
<wf id=”w363”>частиц</wf>.
</div>
```

Здесь указывается вхождение словоформ в предложения. Принцип разложения в пространстве аналогичный: измерение словоформ и измерение предложений.

В функционально-структурной иерархии абзац будет выглядеть вот так:

```
<par type="title" no="6">
<wf id=”w345”>Присовокупление</wf><wf  id=”w346”>III</wf>
</par>
<par type="body" par_no="6" no=”1”>
<wf id=”w347”>6</wf>.<wf id=”w348”>Если</wf><wf id=”w349”>требуется</wf>
<wf id=”w350”>превратить</wf><wf id=”w351”>какое-либо</wf>
<wf id=”w352”>твердое</wf><wf id=”w353”>тело</wf>
```

<wf id="w354">в</wf><wf id="w355">жидкое</wf>,<wf id="w356">то</wf>
<wf id="w357">необходимо</wf><wf id="w358">затруднить</wf>
<wf id="w359">взаимное</wf><wf id="w360">сцепление</wf>
<wf id="w361">отдельных</wf><wf id="w362">его </wf>
<wf id="w363">частиц</wf>.
</par>

В данной разметке для разбиения текста на смысловые блоки присутствует уже несколько подпространств. С помощью тега <par> указывается разбиение текста на блоки, части параграфа и, соответственно, вхождение словоформ в эти блоки (измерения словоформ и блоков). Для блоков указывается, в какой параграф они входят, а также какой семантический тип представляют - заголовок параграфа или тело параграфа. Это уже факт, точка трех измерений: блоки, параграфы, типы блоков. Таким фактом фиксируется одновременно структурная и семантическая информация, знания о тексте. Именно поэтому модель пространства так названа.

Использование в качестве идентификаторов словоформ текста уникальных идентификаторов ресурса (URI) так, как он понимается в схеме RDF (URL: http://www.w3.org/RDF/) и семантическом вебе в целом позволит прозрачно и удобно использовать цифровые тексты различными Интернет системами. В свою очередь оформленные на таком базисе и с использованием модели структурно-семантического пространства остальные знания о текстах также станут понятны любой системе, поддерживающей такое единое описание.

Благодарности

Работа выполнена при поддержке Российского фонда гуманитарных исследований (РГНФ), грант № 07-04-12147в.

Список литературы

Баранов и др. 2007 - *Баранов, В. А.* Интернет портал «Письменное наследие». Формирование сообщества исследователей древних текстов / В. А. Баранов, И. В. Кравцов // Интернет и современное общество : Тр.

X Всерос. объедин. конф. – СПб. : Факультет филол. и искусств СПбГУ, 2007. – С. 57–60.

Кравцов и др. 2007 – *Кравцов, И. В.* Информационная система для работы с коллекциями рукописных исторических документов / И. В. Кравцов, В. О. Филатов // Информационные технологии моделирования и управления. - № 2(36) – Воронеж : Научная книга, 2007. – С. 188–196.

Кравцов 2008 – *Кравцов, И.В.* Моделирование структуры и семантики текста в информационных системах для исследования исторических документов / И. В. Кравцов // Системы управления и информационные технологии. – № 1.1(31) – М. ; Воронеж : Научная книга, 2008. – С. 163–167.

Манускрипт 2008 – *Манускрипт: славянское письменное наследие* [Электронный ресурс]. – 2008. – Режим доступа: http://manuscripts.ru/, свободный. – Загл. с экрана.

О коллекции 2007 – *Манускрипт: Электронная версия произведений М.В. Ломоносова* [Электронный ресурс]. – 2007. – Режим доступа: http://mns.udsu.ru/Lomonosov/Coll_Lmnsv.html, свободный. – Загл. с экрана.

Современные 2006 – *Современные информационные технологии и письменное наследие*: от древних рукописей к электронным текстам : материалы междунар. науч. конф. – Ижевск : Изд-во ИжГТУ, 2006. – 196 с.

RDF 2008 - *Resource Description Framework* (RDF) / W3C Semantic Web Activity [Электронный ресурс]. – 2008. – Режим доступа: http://www.w3.org/RDF, свободный. – Загл. с экрана.

A model for the exchange of knowledge in systems for humanities research

Ignat V. Kravtsov, Kristina A. Bagimova

Petrozavodsk State University, Petrozavodsk, Russia

Izhevsk State Technical University, Izhevsk, Russia

This paper gives a description of a model that produces the structures and semantics of digital texts. This model will form the basis of an XML format for cultural heritage texts produced for the Semantic Web. Such description will allow

us to work with digital texts collections and to make transparent publications and semantic services. This model will also allow knowledge about texts to be exchanged between heterogeneous systems.